Batch Normalization Alleviates the Spectral Bias in Coordinate Networks: Supplemental Material

Zhicheng Cai[1], Hao Zhu[1,*], Qiu Shen[1], Xinran Wang[1,2], Xun Cao[1]
[1]School of Electronic Science and Engineering, Nanjing University, Nanjing, China
[2]Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou, China
Corresponding author: zhuhao_photo@nju.edu.cn

A. Supplementary Derivations

Table 1. Notations of Symbols

Notation	Definition
T	total number of training sample
L	number of hidden layers of the MLP network
$\|\theta\|$	total number of parameters of the MLP network
N^0	the length of each input coordinates vector
N^l	the width of the l-th layer, $l \in [1, L]$
$\phi(\cdot)$	nonlinear activation function
$\vec{x}_i$	the i-th input coordinates, $i \in [1, T]$, $\vec{x}_i \in \mathbb{R}^{N^0}$
$\vec{y}_i$	the signal value corresponding to the i-th input coordinates, $\vec{y}_i \in \mathbb{R}^{N^L}$
$\mathbf{H}^0, \mathbf{X}$	stacked by the T inputs $[\vec{x}_1, \vec{x}_2, ..., \vec{x}_T]$, $\mathbf{X} \in \mathbb{R}^{N^0 \times T}$
$\mathbf{H}^l$	the output of the l-th layer, $\mathbf{H}^l \in \mathbb{R}^{N^l \times T}$
$\mathbf{W}^l$	the weights of the l-th layer, $\mathbf{W}^l \in \mathbb{R}^{N^l \times N^{l-1}}$
$\vec{b}^l$	the bias of the l-th layer, $\vec{b}^l \in \mathbb{R}^{N^l}$
θ	the parameters of the network, $\theta = \{\mathbf{W}^l, \vec{b}^{\,l} \mid l = 1, ..., L\}$
$f(\mathbf{X};\theta),\ \ \mathbf{H}^L$	the final output of the MLP network, $\mathbf{H}^L \in \mathbb{R}^{N^L \times T}$
$f^{BN}(\mathbf{X};\theta)$	the final output of the MLP network with batch normalization
μ	$\mu = \mathbb{E}[\mathbf{H}^L]$, the mean value of $\mathbf{H}^L$
σ	$\sigma = \sqrt{\mathbb{E}[(\mathbf{H}^L)^2] - (\mathbb{E}[\mathbf{H}^L])^2}$, the variance of $\mathbf{H}^L$
$\mathbf{H}$	$\mathbf{H} = \mathbf{H}^L - \mu$, $\mathbf{H} \in \mathbb{R}^{N^L \times T}$
$\mathbf{K}$	the NTK matrix of standard MLP, $\mathbf{K} \in \mathbb{R}^{T \times T}$
$\mathbf{K}^{BN}$	the NTK matrix of MLP with batch normalization, $\mathbf{K}^{BN} \in \mathbb{R}^{T \times T}$
$\nabla_\theta f(\mathbf{X};\theta),\ \ \nabla_\theta \mathbf{H}^L$	deviation to each parameters of the standard MLP output, $\nabla_\theta f(\mathbf{X};\theta) \in \mathbb{R}^{\|\theta\| \times T}$
$\nabla_\theta f^{BN}(\mathbf{X};\theta)$	deviation to each parameters of the BN-based MLP output, $\nabla_\theta f^{BN}(\mathbf{X};\theta) \in \mathbb{R}^{\|\theta\| \times T}$
$\mathbf{D}^l$	the gradients matrix of the l-th layer, $\mathbf{D}^l \in \mathbb{R}^{N^L \times T}$
$p_t^l, q_t^l, p_{st}^l, q_{st}^l$	order parameters of the mean field theory
κ_1, κ_2	two constants corresponding to the order parameters, $\kappa_1 = \frac{1}{L-1}\sum_{l=1}^{L} p_t^{l-1} q_t^l$, $\kappa_2 = \frac{1}{L-1}\sum_{l=1}^{L} p_{st}^{l-1} q_{st}^l$
$\mathbf{M}$	matrix computed by the mean field theory, $\mathbf{M} \in \mathbb{R}^{T \times T}$
$\mathbf{A}$	matrix corresponding to variance division of batch normalization, $\mathbf{A} \in \mathbb{R}^{T \times T}$
$\mathbf{G}$	matrix corresponding to mean subtraction of batch normalization, $\mathbf{G} \in \mathbb{R}^{T \times T}$

[*]This work was supported by the National Natural Science Foundation of China under Grants T2221003, 62071216 and the Leading Technology of Jiangsu Basic Research Plan (BK20192003).

We first introduce all the notations as listed in Table. 1. Note that the derivations are under the hypothesis of infinite (or large enough) network width and sufficient measurements. Thus we can assume that $T = \mathcal{O}(N)$.

A.1. Derivation for Eqn.6

First we rewrite the form of the variance σ in Eqn.5 as:

$$\begin{aligned}
\sigma &= \sqrt{\mathbb{E}[(\mathbf{H}^L)^2 - \mu^2]} = \sqrt{\mathbb{E}[(\mathbf{H}^L)^2 + \mu^2 - 2\mu\mu]} \\
&= \sqrt{\mathbb{E}[(\mathbf{H}^L)^2] + \mathbb{E}[\mu^2] - 2\mathbb{E}[\mathbf{H}^L]\mathbb{E}[\mu]} \\
&= \sqrt{\mathbb{E}[(\mathbf{H}^L)^2 + \mu^2 - 2\mathbf{H}^L\mu]} = \sqrt{\mathbb{E}[(\mathbf{H}^L - \mu)^2]} \\
&= \sqrt{\mathbb{E}[\mathbf{H}^2]} = \sqrt{\frac{\mathbf{H}\mathbf{H}^\top}{T}}
\end{aligned} \tag{A.1}$$

Thus $\nabla_\theta f^{BN}(\mathbf{X};\theta)$ can be calculated as:

$$\begin{aligned}
\nabla_\theta f^{BN}(\mathbf{X};\theta) &= \nabla_\theta \frac{\mathbf{H}^L - \mu}{\sigma} = \nabla_\theta \frac{\mathbf{H}}{\sigma} \\
&= \frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta \sigma \mathbf{H}}{\sigma^2} \\
&= \frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta(\sqrt{\mathbf{H}\mathbf{H}^\top})\mathbf{H}}{\sqrt{T}\sigma^2} \\
&= \frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta(\mathbf{H}\mathbf{H}^\top)\mathbf{H}}{2\sqrt{T}\sigma^2\sqrt{\mathbf{H}\mathbf{H}^\top}} \\
&= \frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{2\nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{2T\sigma^2\sqrt{\frac{\mathbf{H}\mathbf{H}^\top}{T}}} \\
&= \frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{T\sigma^3}
\end{aligned} \tag{A.2}$$

Now we can rewrite the form of $\mathbf{K}^{BN}$ in Eqn.6 as:

$$\begin{aligned}
\mathbf{K}^{BN} &= (\nabla_\theta f^{BN}(\mathbf{X};\theta))^\top \nabla_\theta f^{BN}(\mathbf{X};\theta) \\
&= \left(\frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{T\sigma^3}\right)^\top \left(\frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{T\sigma^3}\right) \\
&= \left(\frac{(\nabla_\theta \mathbf{H})^\top}{\sigma} - \frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top}{T\sigma^3}\right) \left(\frac{\nabla_\theta \mathbf{H}}{\sigma} - \frac{\nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{T\sigma^3}\right) \\
&= \frac{(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{\sigma^2} + \frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}\mathbf{H}^\top\mathbf{H}}{T^2\sigma^6} - 2\frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{T\sigma^4} \\
&= \frac{(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{\sigma^2} + \frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}\sigma^2}{T\sigma^6} - 2\frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{T\sigma^4} \\
&= \frac{(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{\sigma^2} - \frac{\mathbf{H}^\top\mathbf{H}(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{T\sigma^4}
\end{aligned} \tag{A.3}$$

A.2. Order Parameters of Mean Field Theory

Elements in $\nabla_\theta f(\mathbf{X};\theta)$. First we denote $\mathbf{Mat}_{ij}$ is the element in the i-th row and j-th column of $\mathbf{Mat}$. This denotation works for every matrix. Each element of the $|\theta| \times T$ matrix $\nabla_\theta f(\mathbf{X};\theta)$ is computed by the chain rule in a manner similar to that of the back-propagation algorithm, where$|\theta|$ is the number of total parameters, namely, $|\theta| = \sum_{l=1}^{L} N^{l-1}N^l$. The elements of the t-th column of $\nabla_\theta f(\mathbf{X};\theta)$ are the differential coefficient of $f(\vec{x}_t;\theta)$ to all the parameters θ, namely, $\nabla_{\mathbf{W}^l_{ij}} f(\vec{x}_t;\theta)$ for $l = 1, 2, ..., L$, $i = 1, 2, ..., N^{l-1}$, and $j = 1, 2, ..., N^l$. As noted in paper, $\mathbf{W}^l$ is the weight matrix of the l-th fully-connected layer, and randomly initiated with Gaussian distribution, namely, $\mathbf{W}^l_{ij} \sim \mathcal{N}(0, \frac{\sigma^2_{\mathrm{W}}}{N^{l-1}})$. Thus we can compute each element as:

$$
\begin{aligned}
&\nabla_{\mathbf{W}^l_{ij}} f(\vec{x}_t; \theta) = \mathbf{D}^l_{it} \mathbf{H}^{l-1}_{jt}, \quad l = 1, 1, ..., L \\
&\text{where} \quad \mathbf{D}^l_{it} = \phi'(\sum_{n=1}^{N^{l-1}} \mathbf{W}^l_{in} \mathbf{H}^{l-1}_{it} + b^l_i) \sum_{n=1}^{N^{l+1}} \mathbf{W}^{l+1}_{ni} \mathbf{D}^{l+1}_{it}, \quad \mathbf{D}^L_{it} = 1
\end{aligned}
\tag{A.4}
$$

Order parameters of the mean field theory. For the feed-forward signal propagations, we have:

$$
\begin{aligned}
p^l_t &:= \frac{1}{N^l} \sum_{i=1}^{N^l} {\mathbf{H}^l_{it}}^2 \\
p^l_{st} &:= \frac{1}{N^l} \sum_{i=1}^{N^l} \mathbf{H}^l_{is} \mathbf{H}^l_{it}
\end{aligned}
\tag{A.5}
$$

In the context of deep learning, these variables have been utilized to explain the depth to which signals can sufficiently propagate. The variable p^l_{st} is the correlation between the activations for different input samples $\vec{x}_s$ and $\vec{x}_t$ in the l-th layer. Under the large N limit, these variables are given by integration over Gaussian distributions because the pre-activation $\mathbf{W}^l \mathbf{H}^{l-1}$ is a weighted sum of independent random parameters and the central limit theorem is applicable, thus we obtain:

$$
\begin{aligned}
p^{l+1}_t &= \int Du \phi^2 \left(\sqrt{r^{l+1}_t} u \right), \quad r^{l+1}_t = \sigma^2_{\mathrm{W}} p^l_t \\
p^{l+1}_{st} &= I_\phi[r^{l+1}_t, r^{l+1}_{st}], \quad r^{l+1}_{st} = \sigma^2_{\mathrm{W}} p^l_{st}
\end{aligned}
\tag{A.6}
$$

with $p^0_t = 1$ and $p^0_{st} = 0$. The notation Du means integration over the standard Gaussian density, $I_\phi[a, b]$ represents the certain integral:

$$
\begin{aligned}
Du &= du \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \\
I_\phi[a, b] &= \int Dz_1 z_2 \phi \left(\sqrt{a} z_1 \right) \phi \left(\sqrt{a} \left(c z_1 + \sqrt{1 - c^2} z_2 \right) \right), \quad c = \frac{b}{a}
\end{aligned}
\tag{A.7}
$$

Since $p^0_t = 1$ and $p^0_{st} = 0$ are established for all s and t, thus p^l_{st} takes the same value for all $s \neq t$, and so dose p^l_t for all t.

For the back-propagation, we have following variables:

$$
\begin{aligned}
q^l_t &:= \sum_{i=1}^{N^l} {\mathbf{D}^l_{it}}^2 \\
q^l_{st} &:= \sum_{i=1}^{N^l} \mathbf{D}^l_{is} \mathbf{D}^l_{it}
\end{aligned}
\tag{A.8}
$$

Since they are defined not by averages but by the sums, they remain $\mathcal{O}(1)$ because $C = \mathcal{O}(1)$. q_{st} is the correlation of back-propagated signals. To compute these quantities, we have the following recurrence relations ($l = 0, ..., L-1$):

$$
\begin{aligned}
q^l_t &= \sigma^2_{\mathrm{W}} q^{l+1}_t \int Du \left[\phi' \left(\sqrt{r^l_t} u \right) \right]^2 \\
q^l_{st} &= \sigma^2_{\mathrm{W}} q^{l+1}_{st} I_{\phi'}[r^l_t, r^l_{st}]
\end{aligned}
\tag{A.9}
$$

with $q^L_t = q^L_{st} = 1$ because of the linear outputs. Since $q^L_t = q^L_{st} = 1$ are established for all s and t, thus q^l_{st} takes the same value for all $s \neq t$, and so dose q^l_t for all t.

The variables $(p^l_t, q^l_t, p^l_{st}, q^l_{st})$ depend only on the variance parameters σ^2_{W}, not on the unit indices. In that sense, they are referred to as *macroscopic variables*, namely, the order parameters in statistical physics. Furthermore, the recurrence relations for the macroscopic variables simply require L iterations of one- and two-dimensional numerical integrals.

A.3. Derivation for Eqn.7 and Eqn.8

Now we calculate each element $\mathbf{K}_{st}$ in the s-th row and t-th column of the $T \times T$ NTK matrix $\mathbf{K}$:

$$\begin{aligned}
\mathbf{K}_{st} &= ((\nabla_\theta f(\mathbf{X};\theta))^\top \nabla_\theta f(\mathbf{X};\theta))_{st} \\
&= \sum_{l=1}^{L} \sum_{i=1}^{N^l} \sum_{j=1}^{N^{l-1}} \nabla_{\mathbf{W}_{ij}^l} f(\vec{x}_s;\theta) \nabla_{\mathbf{W}_{ij}^l} f(\vec{x}_t;\theta) \\
&= \sum_{l=1}^{L} \sum_{i=1}^{N^l} \sum_{j=1}^{N^{l-1}} \mathbf{D}_{is}^l \mathbf{H}_{js}^{l-1} \mathbf{D}_{it}^l \mathbf{H}_{jt}^{l-1} \\
&= \sum_{l=1}^{L} \sum_{j=1}^{N^{l-1}} \mathbf{H}_{js}^{l-1} \mathbf{H}_{jt}^{l-1} \sum_{i=1}^{N^l} \mathbf{D}_{is}^l \mathbf{D}_{it}^l \\
&= \sum_{l=1}^{L} N^{l-1} \frac{1}{N^{l-1}} \sum_{j=1}^{N^{l-1}} \mathbf{H}_{js}^{l-1} \mathbf{H}_{jt}^{l-1} \sum_{i=1}^{N^l} \mathbf{D}_{is}^l \mathbf{D}_{it}^l
\end{aligned} \tag{A.10}$$

For $s \neq t$, we have:

$$\begin{aligned}
&\frac{1}{N^{l-1}} \sum_{j=1}^{N^{l-1}} \mathbf{H}_{js}^{l-1} \mathbf{H}_{jt}^{l-1} = p_{st}^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \\
&\sum_{i=1}^{N^l} \mathbf{D}_{is}^l \mathbf{D}_{it}^l = q_{st}^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)
\end{aligned} \tag{A.11}$$

Thus for $\mathbf{K}_{st}$:

$$\begin{aligned}
\mathbf{K}_{st} &= \sum_{l=1}^{L} N^{l-1} \left(p_{st}^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) \left(q_{st}^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) \\
&= (L-1)N\kappa_2 + \mathcal{O}(\sqrt{N}) \\
&where, \quad \kappa_2 = \sum_{l=1}^{L} \frac{1}{L-1} p_{st}^{l-1} q_{st}^l
\end{aligned} \tag{A.12}$$

Similarly, for $s = t$, we have:

$$\begin{aligned}
&\frac{1}{N^{l-1}} \sum_{j=1}^{N^{l-1}} \mathbf{H}_{jt}^{l-1} \mathbf{H}_{jt}^{l-1} = p_t^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \\
&\sum_{i=1}^{N^l} \mathbf{D}_{it}^l \mathbf{D}_{it}^l = q_t^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)
\end{aligned} \tag{A.13}$$

Thus for $\mathbf{K}_{tt}$:

$$\begin{aligned}
\mathbf{K}_{tt} &= \sum_{l=1}^{L} N^{l-1} \left(p_t^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) \left(q_t^l + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) \\
&= (L-1)N\kappa_1 + \mathcal{O}(\sqrt{N}) \\
&where, \quad \kappa_1 = \sum_{l=1}^{L} \frac{1}{L-1} p_t^{l-1} q_t^l
\end{aligned} \tag{A.14}$$

Thus the $T \times T$ NTK matrix of the coordinate network $\mathbf{K} = (\nabla_\theta \mathbf{H}^L)^\top \nabla_\theta \mathbf{H}^L$ can be calculated and re-written in the explicit numerical form as:

$$\mathbf{K}_{ij} = \begin{cases} (L-1)N\kappa_1 + \mathcal{O}(\sqrt{N}), & for\ \ i = j \\ (L-1)N\kappa_2 + \mathcal{O}(\sqrt{N}), & for\ \ i \neq j \end{cases} \tag{A.15}$$

namely:

$$\mathbf{K} = (L-1)N\mathbf{M} + \mathcal{O}(\sqrt{N}) \tag{A.16}$$

where $\mathbf{M}$ is a $T \times T$ matrix:

$$\mathbf{M} = \begin{bmatrix} \kappa_1 & \kappa_2 & \cdots & \kappa_2 \\ \kappa_2 & \kappa_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \kappa_2 \\ \kappa_2 & \cdots & \kappa_2 & \kappa_1 \end{bmatrix}$$
$$\kappa_1 = \frac{1}{L-1}\sum_{l=1}^{L} p_t^{l-1} q_t^l, \quad \kappa_2 = \frac{1}{L-1}\sum_{l=1}^{L} p_{st}^{l-1} q_{st}^l \tag{A.17}$$

κ_1 and κ_2 are two constant values corresponding to the order parameters of the mean field theory $p_t^l, q_t^l, p_{st}^l, q_{st}^l$.

A.4. Derivation for Eqn.9

$$\begin{aligned} \mathbf{K}^{BN} &= \frac{(\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{\sigma^2} - \frac{\mathbf{H}^\top \mathbf{H} (\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}}{T\sigma^4} \\ &= \left(\frac{\mathbf{I}}{\sigma^2} - \frac{\mathbf{H}^\top \mathbf{H}}{T\sigma^4} \right) ((\nabla_\theta \mathbf{H})^\top \nabla_\theta \mathbf{H}) \\ \because \mathbf{H} &= \mathbf{H}^L - \mu \\ \therefore \nabla_\theta \mathbf{H} &= \nabla_\theta \mathbf{H}^L - \nabla_\theta \mu \\ \therefore \mathbf{K}^{BN} &= \left(\frac{\mathbf{I}}{\sigma^2} - \frac{\mathbf{H}^\top \mathbf{H}}{T\sigma^4} \right) (\nabla_\theta \mathbf{H}^L - \nabla_\theta \mu)^\top (\nabla_\theta \mathbf{H}^L - \nabla_\theta \mu) \\ &= \mathbf{A} (\nabla_\theta \mathbf{H}^L - \nabla_\theta \mu)^\top (\nabla_\theta \mathbf{H}^L - \nabla_\theta \mu) \\ \text{where } \ \mathbf{A} &= \frac{1}{\sigma^2} \left(\mathbf{I} - \frac{1}{T} \frac{\mathbf{H}^\top \mathbf{H}}{\sigma^2} \right) \end{aligned} \tag{A.18}$$

A.5. Derivation for Eqn.11 and Eqn.12

We have obtained in paper that $\mathbf{K}^{BN} = (L-1)N\mathbf{A}\mathbf{M}^{BN} + \mathcal{O}(\sqrt{N})$, and $\mathbf{M}^{BN} = \mathbf{G}^\top \mathbf{M} \mathbf{G}$. Thus we first derive the numerical value of each element in $\mathbf{M}^{BN}$, then we deduce $\mathbf{A}\mathbf{M}^{BN}$.

Derivation for Eqn.11. To clarify the derivation, we first calculate $\mathbf{R} = \mathbf{G}^\top\mathbf{M}$, then calculate $\mathbf{M}^{BN} = \mathbf{R}\mathbf{G}$:

$$\because \mathbf{G} = \begin{bmatrix} \frac{T-1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & \frac{T-1}{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{T} \\ -\frac{1}{T} & \cdots & -\frac{1}{T} & \frac{T-1}{T} \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} \kappa_1 & \kappa_2 & \cdots & \kappa_2 \\ \kappa_2 & \kappa_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \kappa_2 \\ \kappa_2 & \cdots & \kappa_2 & \kappa_1 \end{bmatrix}$$

$$\therefore \mathbf{G}_{ij} = \begin{cases} \frac{T-1}{T}, & for\ \ i=j \\ -\frac{1}{T}, & for\ \ i \neq j \end{cases}, \quad \mathbf{M}_{ij} = \begin{cases} \kappa_1, & for\ \ i=j \\ \kappa_2, & for\ \ i \neq j \end{cases}$$

$$\therefore \mathbf{R}_{ij} = \begin{cases} \frac{T-1}{T}\kappa_1 - (T-1)\frac{1}{T}\kappa_2, & for\ \ i=j \\ -\frac{1}{T}\kappa_1 + \frac{T-1}{T}\kappa_2 - (T-2)\frac{1}{T}\kappa_2, & for\ \ i \neq j \end{cases}$$

$$= \begin{cases} \frac{T-1}{T}(\kappa_1 - \kappa_2), & for\ \ i=j \\ \frac{1}{T}(\kappa_2 - \kappa_1), & for\ \ i \neq j \end{cases}$$

$$\therefore \mathbf{M}^{BN}_{ij} = \begin{cases} \frac{T-1}{T}\frac{T-1}{T}(\kappa_1 - \kappa_2) - (T-1)\frac{1}{T}\frac{1}{T}(\kappa_2 - \kappa_1), & for\ \ i=j \\ \frac{T-1}{T}\frac{1}{T}(\kappa_2 - \kappa_1) - \frac{1}{T}\frac{T-1}{T}(\kappa_1 - \kappa_2) - (T-2)\frac{1}{T}\frac{1}{T}(\kappa_2 - \kappa_1), & for\ \ i \neq j \end{cases} \tag{A.19}$$

$$= \begin{cases} \frac{(T-1)^2 + (T-1)}{T^2}(\kappa_1 - \kappa_2), & for\ \ i=j \\ \frac{(T-1)+(T-1)-(T-2)}{T^2}(\kappa_2 - \kappa_1), & for\ \ i \neq j \end{cases}$$

$$= \begin{cases} \frac{T(T-1)}{T^2}(\kappa_1 - \kappa_2), & for\ \ i=j \\ \frac{T}{T^2}(\kappa_2 - \kappa_1), & for\ \ i \neq j \end{cases}$$

$$= \begin{cases} \frac{T-1}{T}(\kappa_1 - \kappa_2), & for\ \ i=j \\ \frac{1}{T}(\kappa_2 - \kappa_1), & for\ \ i \neq j \end{cases}$$

$$\therefore \mathbf{M} = \begin{bmatrix} \frac{(T-1)}{T}(\kappa_1 - \kappa_2) & \frac{1}{T}(\kappa_2 - \kappa_1) & \cdots & \frac{1}{T}(\kappa_2 - \kappa_1) \\ \frac{1}{T}(\kappa_2 - \kappa_1) & \frac{(T-1)}{T}(\kappa_1 - \kappa_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{T}(\kappa_1 - \kappa_2) \\ \frac{1}{T}(\kappa_1 - \kappa_2) & \cdots & \frac{1}{T}(\kappa_1 - \kappa_2) & \frac{(T-1)}{T}(\kappa_1 - \kappa_2) \end{bmatrix}$$

Derivation for Eqn.12. Now we can deduce the formulation of $\mathbf{A}\mathbf{M}^{BN}$:

$$\begin{aligned} \mathbf{A}\mathbf{M}^{BN} &= \frac{\kappa_1 - \kappa_2}{\sigma^2}(\mathbf{I} - \frac{1}{T}\frac{\mathbf{H}^\top\mathbf{H}}{\sigma^2}) \begin{bmatrix} \frac{T-1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & \frac{T-1}{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{T} \\ -\frac{1}{T} & \cdots & -\frac{1}{T} & \frac{T-1}{T} \end{bmatrix} \\ &= \frac{\kappa_1 - \kappa_2}{\sigma^2}\left(\begin{bmatrix} \frac{T-1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & \frac{T-1}{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{T} \\ -\frac{1}{T} & \cdots & -\frac{1}{T} & \frac{T-1}{T} \end{bmatrix} - \frac{1}{T\sigma^2}\mathbf{H}^\top\mathbf{H} \begin{bmatrix} \frac{T-1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & \frac{T-1}{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{T} \\ -\frac{1}{T} & \cdots & -\frac{1}{T} & \frac{T-1}{T} \end{bmatrix} \right) \\ &\approx \frac{\kappa_1 - \kappa_2}{\sigma^2}(\mathbf{I} - \frac{1}{T\sigma^2}\mathbf{H}^\top\mathbf{H}\frac{T-1}{T}\mathbf{I}) \\ &= \frac{\kappa_1 - \kappa_2}{\sigma^2}(\mathbf{I} - \frac{1}{T\sigma^2}(1 - \frac{1}{T})\mathbf{H}^\top\mathbf{H}) \\ &\approx \frac{\kappa_1 - \kappa_2}{\sigma^2}\mathbf{I} \end{aligned} \tag{A.20}$$

$$\text{thus:} \quad \mathbf{K}^{BN} = (L-1)N\mathbf{A}\mathbf{M}^{BN} + \mathcal{O}(\sqrt{N}) \approx (L-1)N\frac{\kappa_1 - \kappa_2}{\sigma^2}\mathbf{I} + \mathcal{O}(\sqrt{N})$$

A.6. Derivation for Eqn.14

Here we calculate the statistic characteristics (namely, mean value m_λ, second momentum s_λ variance v_λ and the maximum eigenvalue λ_{max}) of the NTKs' eigenvalues of standard MLP and BN-based MLP in detail. According to the matrix theory, these values could be calculated or estimated by,

$$
\begin{aligned}
& m_\lambda = \frac{1}{T} Trace(\mathbf{K}) = \frac{1}{T}\sum_{i=1}^{T} K_{ii} \\
& s_\lambda = \frac{1}{T}\sum_{i=1}^{T}\lambda_i^2 = \frac{1}{T} Trace(\mathbf{K}\mathbf{K}^\top) = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{K}\mathbf{K}^\top)_{ii} \\
& v_\lambda = s_\lambda - m_\lambda^2 \\
& \frac{s_\lambda}{m_\lambda} = \frac{\sum_{i=1}^{T}\lambda_i^2}{\sum_{i=1}^{T}\lambda_i} \le \lambda_{max} \le \sqrt{\sum_{i=1}^{T}\lambda_i^2} = \sqrt{T s_\lambda}
\end{aligned}
\tag{A.21}
$$

Standard MLP.

For the original coordinate network, the NTK can be explicitly written as:

$$
\begin{aligned}
& \because \mathbf{K} = (L-1)N\mathbf{M} + \mathcal{O}(\sqrt{N}) \\
& \therefore \mathbf{K}_{ij} = \begin{cases} (L-1)N\kappa_1 + \mathcal{O}(\sqrt{N}), & for\ \ i = j \\ (L-1)N\kappa_2 + \mathcal{O}(\sqrt{N}), & for\ \ i \neq j \end{cases}
\end{aligned}
\tag{A.22}
$$

Thus the mean value can be calculated as:

$$
m_\lambda = \frac{(L{-}1)NT\kappa_1 + \mathcal{O}(\sqrt{N})T}{T} \sim \mathcal{O}(N) \tag{A.23}
$$

To calculate the second momentum, we first calculate the $\mathbf{S} = \mathbf{K}\mathbf{K}^\top$ as(for simplicity, we denote $(L-1)\kappa_1, (L-1)\kappa_2$ as $\tilde{\kappa}_1, \tilde{\kappa}_2$):

$$
\begin{aligned}
\mathbf{S}_{ij} &= \begin{cases} (N\tilde{\kappa}_1 + \mathcal{O}(\sqrt{N}))^2 + (T-1)(N\tilde{\kappa}_2 + \mathcal{O}(\sqrt{N}))^2, & for\ \ i = j \\ 2(N\tilde{\kappa}_1 + \mathcal{O}(\sqrt{N}))(N\tilde{\kappa}_2 + \mathcal{O}(\sqrt{N})) + (T-2)(N\tilde{\kappa}_2 + \mathcal{O}(\sqrt{N})), & for\ \ i \neq j \end{cases} \\
&= \begin{cases} N^2((T-1)\tilde{\kappa}_2^2 + \tilde{\kappa}_1^2) + 2N\mathcal{O}(\sqrt{N})((T-1)\tilde{\kappa}_2 + \tilde{\kappa}_1) + \mathcal{O}(N)T, & for\ \ i = j \\ N^2((T-2)\tilde{\kappa}_2^2 + 2\tilde{\kappa}_1\tilde{\kappa}_2) + 2N\mathcal{O}(\sqrt{N})((T-1)\tilde{\kappa}_2 + \tilde{\kappa}_1) + \mathcal{O}(N)T, & for\ \ i \neq j \end{cases}
\end{aligned}
\tag{A.24}
$$

Thus the second momentum can be calculated as:

$$
s_\lambda = \frac{1}{T}\sum_{i=1}^{T} S_{ii} = N^2((T-1)\tilde{\kappa}_2^2 + \tilde{\kappa}_1^2) + 2N\mathcal{O}(\sqrt{N})((T-1)\tilde{\kappa}_2 + \tilde{\kappa}_1) + \mathcal{O}(N)T \sim \mathcal{O}(N^3) \tag{A.25}
$$

Now we can accordingly deduce the variance and maximum eigenvalue as:

$$
\begin{aligned}
& v_\lambda \sim \mathcal{O}(N^3) - \mathcal{O}(N^2) \sim \mathcal{O}(N^3) \\
& \mathcal{O}(N^2) \le \lambda_{max} \le \mathcal{O}(N^2),\ \ \lambda_{max} \sim \mathcal{O}(N^2)
\end{aligned}
\tag{A.26}
$$

BN-based MLP.

For the coordinate network with batch normalization, the NTK $\mathbf{K}^{BN}$ can be explicitly written as:

$$
\begin{aligned}
& \because \mathbf{K}^{BN} \approx (L-1)N\frac{\kappa_1 - \kappa_2}{\sigma^2}\mathbf{I} + \mathcal{O}(\sqrt{N}) \\
& \therefore \mathbf{K}_{ij}^{BN} \approx \begin{cases} (L-1)N\frac{\kappa_1 - \kappa_2}{\sigma^2} + \mathcal{O}(\sqrt{N}), & for\ \ i = j \\ \mathcal{O}(\sqrt{N}), & for\ \ i \neq j \end{cases}
\end{aligned}
\tag{A.27}
$$

Value type	Standard MLP.	BN-based MLP.
m_{λ}	$\mathcal{O}(N)$	$\mathcal{O}(N)$
v_{λ}	$\mathcal{O}(N^3)$	$\mathcal{O}(N^{\frac{3}{2}})$
λ_{max}	$\mathcal{O}(N^2)$	$[\mathcal{O}(N), \mathcal{O}(N^{\frac{3}{2}})]$

Table 2. Comparisons of statistical characteristics of eigenvalues in NTKs without and with batch normalization.

Thus the mean value can be calculated as:

$$m_{\lambda} = \frac{(L-1)}{\sigma^2}\frac{NT(\kappa_1 - \kappa_2)}{T} + \frac{\mathcal{O}(\sqrt{N})T}{T} \sim \mathcal{O}(N) \tag{A.28}$$

To calculate the second momentum, we first calculate the $\mathbf{S}^{BN} = \mathbf{K}^{BN}(\mathbf{K}^{BN})^{\top}$ as(for simplicity, we denote $\frac{(L-1)}{\sigma^2}(\kappa_1 - \kappa_2)$ as κ):

$$\begin{aligned}\mathbf{S}_{ij}^{BN} &= \begin{cases} (N\kappa + \mathcal{O}(\sqrt{N}))^2 + (T-1)(\mathcal{O}(\sqrt{N}))^2, & for \ \ i = j \\ 2\mathcal{O}(\sqrt{N})(N\kappa + \mathcal{O}(\sqrt{N})) + (T-2)(\mathcal{O}(\sqrt{N}))^2, & for \ \ i \neq j \end{cases} \\ &= \begin{cases} N^2\kappa^2 + T\mathcal{O}(N) + 2N\mathcal{O}(\sqrt{N})\kappa, & for \ \ i = j \\ 2N\mathcal{O}(\sqrt{N})\kappa + T\mathcal{O}(N), & for \ \ i \neq j \end{cases}\end{aligned} \tag{A.29}$$

Thus the second momentum can be calculated as:

$$s_{\lambda} = N^2\kappa^2 + T\mathcal{O}(N) + 2N\mathcal{O}(\sqrt{N})\kappa \sim \mathcal{O}(N^2) + \mathcal{O}(N^{\frac{3}{2}}) \tag{A.30}$$

Now we can accordingly deduce the variance and maximum eigenvalue as:

$$\begin{aligned}& v_{\lambda} \sim \mathcal{O}(N^2) + \mathcal{O}(N^{\frac{3}{2}}) - \mathcal{O}(N^2) \sim \mathcal{O}(N^{\frac{3}{2}}) \\ & \mathcal{O}(N) \leq \lambda_{max} \leq \mathcal{O}(N^{\frac{3}{2}})\end{aligned} \tag{A.31}$$

We summary the statistical characteristics of NTKs' eigenvalues in Table 2.